Towards Intelligent Open Data Platforms

Discovering Relatedness in Datasets

Oladipupo A. Sennaike[1], Mohammad Waqar[2], Edobor Osagie[2], Islam Hassan[2], Arkadiusz Stasiewicz[2], Lukasz Porwol[2], Adegboyega Ojo[2*]
Department of Computer Science, University of Lagos, Nigeria[1]
osennaike@unilag.edu.ng
Insight Centre for Data Analytics, National University of Ireland, Galway (NUIG), Rep. of Ireland[2]
{name.surname}@insight-centre.org

***Abstract*—Open data platforms are central to the management and exploitation of data ecosystems. While existing platforms provide basic search capabilities and features for filtering search results, none of the existing platforms provide recommendations on related datasets. Knowledge of dataset relatedness is critical for determining datasets that can be mashed-up or integrated for the purpose of analysis and creation of data-driven services. When considering data platforms, such as data.gov with over 193,000 datasets or data.gv.uk with over 40,000 datasets, specifying dataset relatedness relationship manually is infeasible. In this paper, we approach the problem of discovering relatedness in datasets by employing the Kohonen Self Organsing Map (SOM) algorithm to analyze the metadata extracted from the Data Catalogue maintained on a platform. Our results show that this approach is very effective in discovering relatedness relationships among datasets. Findings also reveal that our approach could uncover interesting and valuable connections among domains of the datasets which could be further exploited for designing smarter data-driven services.**

Keywords—Semantic relatedness of datasets; data recommendation; open data platforms; e-government

I. INTRODUCTION

Open data platforms are central to data ecosystems. These data infrastructures mediate public access to the increasingly available open government and public data. In addition to providing access to available data, open data platforms enable organizations to manage their data catalogues, publish, explore, analyse and share their datasets. Currently, there are over ten known open data platforms including CKAN, DKAN, Socrata, PublishMyData, Information Workbench, Enigma, Junar, OpenDataSoft, Callimachus, DataTank and Semantic Media Wiki [1].

Following the proliferation of these portals and datasets published on them, attention has shifted to assessment and evaluation of these portals and the underlying platforms in terms of their features and affordances. For instance authors of [2] evaluated 12 platforms used by National Statistical Offices to determine the degree to which these platforms support structural metadata, Online Analytical Processing (OLAP) hypercubes, Data endpoints, online analysis and visualisation as well as user experience and customisation. The degree to which open data platforms support data transparency features and extendibility was investigated in [1]. The authors of [3] also examined the ease of installation as well as the performance of the platforms. In addition, questions have been raised on the actual use and usability of existing open data platforms [4].

While existing platforms provide basic search capabilities and features for filtering search results, none of the existing platforms provide recommendations on related datasets. Knowledge of dataset relatedness is critical for determining datasets that can be mashed-up or integrated for the purpose of analysis and creation of data-driven services. When considering data platforms with large number of publishers and datasets such as data.gov with over 193,000 datasets or data.gov.uk with over 40,000 datasets, specifying dataset relatedness relationship manually is infeasible.

Over the years, many techniques have been developed for determining semantic relatedness of documents usually modelled as term vectors. Roughly these techniques either use resources such as Wikipedia or WordNet for computing their semantic measures or employ topics or similar structures generated from a corpus of interest. For example, the Explicit Semantic Analysis technique described in [5], [6] employ Wikipedia for computing semantic relatedness. Works described in [7], [8], [9] and [10] are based on WordNet. Techniques such as the Latent Dirichlet Allocation (LDA) [11] used for topic modelling could also be employed for establishing sematic relatedness of documents. In this case documents sharing the same topics are semantically related.

Self-organizing Map (SOM), an unsupervised Neural Network model can also be used for determining semantic relatedness in texts [12]. Specifically, SOM projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data [13]. The input space is usually multidimensional while the grid represents spatially organized internal representations of various features of the input data and their abstractions [14]. The nodes in SOM grids represent some form of classes for the rows of the dataset. The spatial features of the SOM maps provides conceptually richer representation of categories than other methods that provide only scalar values as measures of semantic distances or relatedness. A similar attempt at extending the LDA algorithm with visual representation of topics (LDAvis) was presented in [15].

In this paper, we approach the problem of discovering relatedness in datasets by employing the Kohonen Self Organising Map (SOM) algorithm to analyse the metadata extracted from the data catalogue of Dublin City Open Data

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

Platform – DubLinked. The catalogue comprises 255 open datasets published by several government departments of the Dublin City Council. The input dataset to our SOM model comprised term vectors obtained from the analysis of selected metadata elements for each of the 255 datasets. The resulting SOM map was employed for deducing implicit semantic relatedness among the datasets.

The rest of the paper is organized as: Section 2 describes some of the features and affordances of contemporary open data platforms with some emphasis on search and dataset recommendation features. Our approach to computing dataset relatedness using the SOM approach is presented in Section 3. Results are presented in Section 4 and discussions in Section 5. Finally, concluding remarks are presented in Section 6.

II. Open Data Platforms

Open Data Platforms (ODP) are technological infrastructure comprising of a software ecosystem that supports different end-user interactions with open data including search, discovery of related datasets, publishing, metadata management, sharing, analysis and visualization [1]. A key purpose of open data platforms is to promote access to government data and encourage development of creative tools and applications to engage and serve the wider community [4].

There are at least three major categories of users for open data platforms. The first category is the general public with basic data literacy and low technical skills. The second category expert users comprising data scientists/engineers and software developers capable of carrying out advanced technical work such as analytics and use of application programming interfaces to access data endpoints or catalogues. The third category is the publishers (usually government agencies or entities) responsible for publishing datasets. Effectively supporting ordinary users remains a challenge on the various open data platforms.

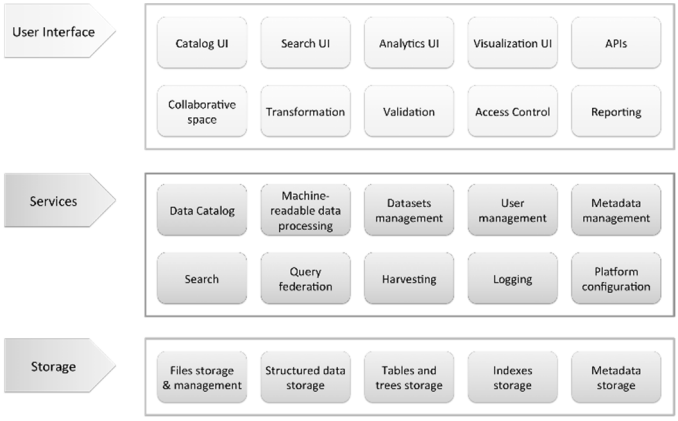

Fig. 1. Generic Architecture of Open Data Portals

According to [16], ODPs provide a number of data management services, data browsing and content management services. However, a detailed study by the authors documented in [1] show that ODPs have at roughly three main sub-systems or layers – User Interface of front end, Services layer and Storage Layer (see Fig. 1).

The user interface layer provides information on available datasets and the search services among others. End-users interact with this layer also to visualize and analyse datasets. Technical users like developers access the catalogue and associated datasets through APIs provided.

The services layer is the heart of the platform as it provides all services that are accessible through the user interface layer. The last layer deals with storage and management of datasets including indexes and specifically storage metadata, as well as contents. Some of the other findings from the exploration of instances of 11 ODPs listed in Table 1 reveals the following affordances and features:

- *Support different metadata schema and data file formats:* Metadata is the information which defines the nature and content of datasets. It includes information such as: title, description, publisher, date of publication, keywords and associated geographical locations. Well defined metadata is essential for indexing, understanding datasets and recommendation. Most platforms allow platform owners in collaboration with publishers to define (or adopt existing) metadata standards. These platforms also allow dataset resources in different data file formats.
- *Allows organizations to publish their datasets***:** Is the process of uploading new datasets, managing old datasets with associated metadata.
- *Provide search facility for datasets:* Most platforms allow keyword-based search on metadata associated with the dataset. Very few platforms such as Enigma allows record- or content-level search.

TABLE I. Open Data Platforms Reviewed by Authors

Platform	URL
CKAN	http://ckan.org/
DKAN	http://nucivic.com/dkan/
Socrata	http://www.socrata.com/
PublishMyData	http://www.swirrl.com/publishmydata
Information Workbench	http://www.fluidops.com/en/portfolio/information_workbench/
Enigma	http://enigma.io/
Junar	http://www.junar.com/
OpenDataSoft(ODS)	http://www.opendatasoft.com/
Callimachus	http://www.callimachus.com/
DataTank	http://www.datatank.co.uk/
Semantic Media Wiki	https://semantic-mediawiki.org/wiki/Semantic_MediaWiki

- *Enable sharing of information on dataset on social media channels*: Most platforms allow users to share information about datasets on social media platforms like Facebook. Platforms could also be configured to share contents when events like publication or download of datasets occurs. Emerging platforms offer users features for dataset rating and feedback, collaborative curation of datasets, voting on dataset requests, reward system and gamification [17][18].
- *Enable federation and harvesting of datasets from different data sources and plaforms*: Federation gives a seamless experience across different platform instance

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

by replicating data across different instances of the platform.

- *Provide data analysis and visualisation tools*: This feature allows the user to explore, analyze, query and summarize datasets. It aims at sense-making and understanding of datasets. However, most platforms currently provide only basic charting features.
- *Allow extensions to core features*: A good number of platforms (in particular CKAN) allow developers to extend available features for instance as plugins.
- *Support personalization*: Existing platforms allow users to customize the look-and-feel of the platform according to their desire.
- *Support programmatic access to data resources*: A good number of platforms provide API for developers to programmtically access catalogs and datasets they manage as resources for external applications.

However, we found that recommendation features were largely not provided by the current generation of open data platforms. For instance, when search results are listed, there are no features that suggest related datasets (see Table 2 for summary of search features on platforms). Given that most practical data use scenarios involve the use of more than one dataset [4], recommendations of related datasets is a very desirable affordance. Our work attempts to address this important shortcoming of current generation of ODPs.

TABLE II. OPEN DATA PLATFORMS SEARCH FEATURES

Platforms	Search feature
CKAN	CKAN provides both search UI and search API on metadata fields with support for filtering.
DKAN	DKAN provides search UI and allows filtering on metadata fields
Socrata	Provides search service over the dataset description and allows filtering.
PublishMyData	Providing limited keyword search on dataset catalogue.
Information Workbench	Provides no user interface or API for searching
Enigma	Provides powerful search user interface and API for search at record level.
Junar	Limited search service
OpenDataSoft	Allow keyword based search and provides API for searching
Callimachus	Not supported
Datatank	Limited filtering by dataset name
Semantic Media Wiki	Free text search over data and allows limited filtering

III. METHODOLOGY

This section describes in details our SOM-based approach to computing of semantic relatedness of datasets. We introduce the SOM model, description of our training datasets, our process for model selection and how we evaluated our model.

A. Research Objectives

Our goal in this work is to determine the implicit semantic relatedness of datasets published as part of a catalogue using Self organizing map as basis for dataset recommendations during search or similar operations on the open data platform.

B. Self-organizing Maps

The Self Organising Map (SOM) is an unsupervised artificial neural network that projects high dimensional data unto a low (usually two) dimensional space while preserving topological order. Order preservation implies that related data are close on the resulting map. The map consists of an array of units or nodes arranged in a regular rectangular or hexagonal grid. Each node has an associated *n*-dimensional model vector $\boldsymbol{m}_k = [m_{k1}, \dots \dots, m_{kn}] \in \mathbb{R}^n$ that approximates the set of input data, where *n* is the dimension of the input space.

The SOM is a competitive, winner take all neural network. During training, a data item $\boldsymbol{x} = [x_1, \dots \dots, x_n] \in \mathbb{R}^n$ is presented to the nodes in parallel. The nodes compete and the node with the best matching model vector, based on a metric, emerges as the winner. The Euclidean distance metric is usually used. The model vectors of the best matching unit and its neighbours are then adjusted so that they are closer to the input data.

$$\boldsymbol{m}_k(t+1) = \boldsymbol{m}_k(t) + h_{c(x),k}(t)[\boldsymbol{x}(t) - \boldsymbol{m}_k(t)], \quad (1)$$

Where, *t* is a time step and $h_{c(x),k}(t)$ is the neighborhood function [19], and

$$c(x) = \arg\min_k \{\|\boldsymbol{x} - \boldsymbol{m}_k\|\}, \quad (2)$$

is the best matching unit.

The neighbourhood function is usually a Gaussian function

$$h_{c(x),k}(t) = \alpha(t) \exp\left(-\frac{\|r_k - r_{c(x)}\|^2}{2\sigma^2(t)}\right), \quad (3)$$

Where, $0 < \alpha(t) < 1$ is the learning-rate, $r_k \in \mathbb{R}^2$ and $r_{c(x)} \in \mathbb{R}^2$ are vectorial locations on the display grid, and $\sigma(t)$ corresponds to the width of the neighborhood function. Both $\alpha(t)$ and $\sigma(t)$ decrease monotonically with the time steps.

In our work, we employ the Batch Map; a computationally effective version of the original SOM [19].

C. Description of Dataset

The data used for training the model is extracted from the Dublin City Council (DubLinked)[1] instance of the CKAN platform using the REST API; CKAN API are used to get the list of 255 available datasets and the associated metadata; the content and field names from tabular data are extracted using DataStore[2] extension for CKAN; metadata and content are integrated; the results are passed to named entity recognition (NER) library to extract entities like person, organization and location from the metadata and the content. The resulting features are listed in TABLE III. Elements of the extracted data include title, package id, organization, theme, notes, tags, resource id and resource fields.

[1] http://dublinked.ie/
[2] http://docs.ckan.org/en/latest/maintaining/datastore.html

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

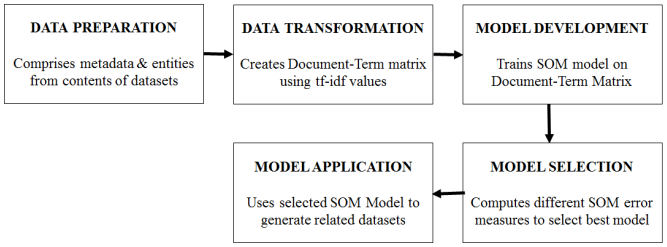

Fig. 2. SOM Model Development Process

Following the data extraction or preparation step, the dataset was transformed into a Document Term matrix using the term frequency–inverse document frequency (tf-idf) metrics (see Fig. 2). The transformation produced a 255 by 1241 matrix (i.e. for the 255 datasets). The document-term matrix was served as input to an SOM.

TABLE III. FEATURES OF PREPARED DATASET

Feature	Description	Source
Title	Title of the dataset	Metadata
Organization	Organization which published the dataset	Metadata
Theme	Theme or category of the dataset	Metadata
Notes	Textual description of the dataset	Metadata
Tag	A set tag that are related to the dataset.	Metadata
Resource Fields	Fields name extracted from tabular resource associated with the dataset	Content
Location	Location related entities extracted from the content and metadata	NER
Person	Person type entities extracted from the content and metadata	NER
Organization	Organization type entities extracted from the content and metadata	NER

D. Model Development and Selection

Traditionally, selection of a map size is usually based on heuristics. For our work, we base the size of the map on a measure. A number of organisation measures [20] exist for the SOM algorithm. These include Topological Error, Quantisation Error, Inversion Measure, Sammon Measure, Spearman Coefficient, Minimal Path Length, Minimal wiring, Minimal Distortion and Inverted Minimal Distortion [21] [22]. For our work, we consider the topological error and quantisation error in selecting the size of our map.

Topological error is the proportion of input samples for which the first and second best matching units are not neighbours on the map. Thus, a large topological error indicates that the map does not preserve the topology of the high dimensional input. For our relatedness objective, this measure is very important as related datasets must be close on the resulting map. We further extended the topographic error by introducing a neighbourhood radius. Thus a neighbourhood radius of 2 will encompass nodes that are two degrees away from the node under consideration.

The quantization error is the average distance between the input data and the model vector of its best matching unit. Thus, this indicates how well a map represents the training data. Small quantisation error is desirable as it indicates that the map matches the input data.

Below is a table showing the average values of the topographic errors (with radius 1 and 2) and the quantisation error for different map sizes (Table 4).

TABLE IV. ERRORS FOR DIFFERENT SOM MAP SIZES

Map Size	Topographic Error	Topographic Error (radius=2)	Quantisation Error
5 by 10	0.006666667	0.005490196	0.876504558
10 by 10	0.000392157	0.000392157	0.829669328
10 by 15	0.000392157	0.000392157	0.788945684
15 by 15	0.003529412	0.003137255	0.743332075
15 by 20	0.000392157	0	0.698631404
20 by 20	0.000784314	0.000392157	0.658406129
20 by 25	0.001960784	0.000784314	0.614880661
30 by 35	0.003137255	0	0.454486411
40 by 50	0.010980392	0.003137255	0.326628873
50 by 60	0.017254902	0.001176471	0.263181763
60 by 60	0.030588235	0.008627451	0.253023027
60 by 70	0.058431373	0.012156863	0.228825969
70 by 70	0.078039216	0.028627451	0.231835478

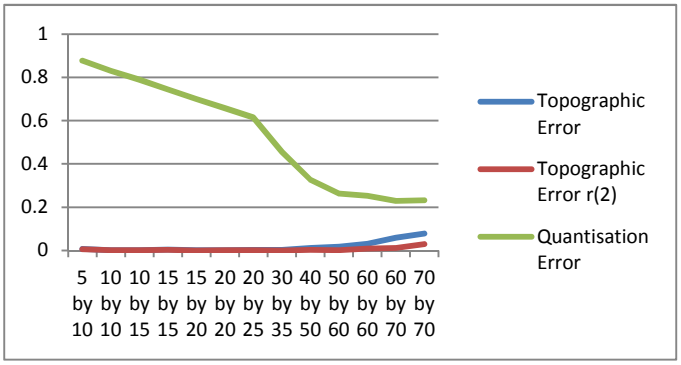

Fig. 3. Topographic & Quantization Errors

From the graph (Fig.3) we decided to choose a map of 20 by 20 nodes as the topographic errors were still at a minimum.

E. Validation and Evaluation

In order to validate the model, domain experts examined the categories as discovered by the map. A category will be defined by the data items whose best matching unit is being considered, and those within a specified radius. In all cases, the domain experts were able to infer the concept being addressed by each category, even when the radius is increased reasonably. See discussion for examples.

IV. RESULTS

A. Resulting Topographic Map of Datasets

The generated topographic map is visualized below (Fig. 4). For better map readability, the datasets were labelled serially which are shown on the map. For nodes that have multiple data items, these numbers overlap and the map only shows one.

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

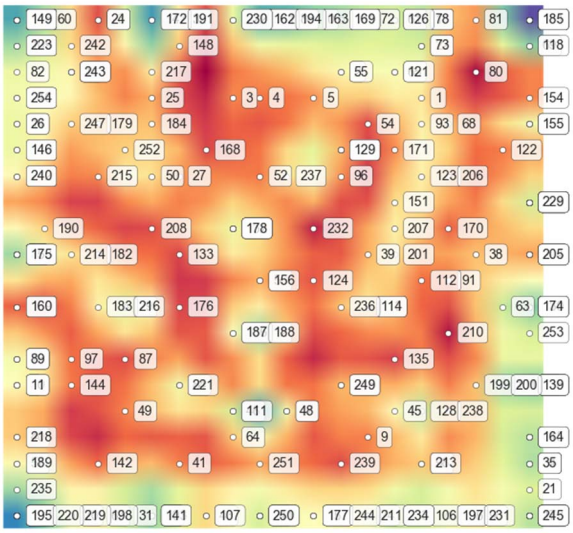

Fig. 4. Resulting 20 by 20 SOM Map.

B. *Distribution of Dataset per Nodes*

The heat map in Fig. 5 shows the distribution of the data items on the map. The map shows relatively good spread of datasets across nodes with relatively more datasets appearing to cluster at the boundaries. Fig. 6 shows that the most nodes have around 10 datasets.

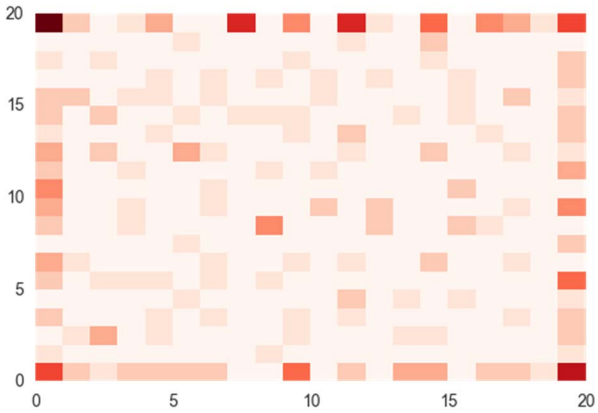

Fig. 5. Heat map for selected SOM Map (20 by 20).

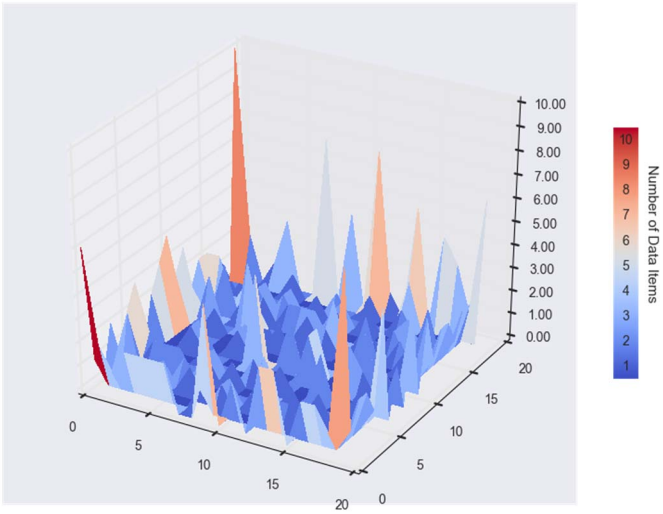

Fig. 6. Distribution of datasets per Nodes on SOM Map.

C. *Top Terms for Nodes and their Neigbours*

The top terms for a node somewhat define the node. The node with the top data items has 10 data items. The common terms in this node are "art", "heritage", "culture" and "dublink". These terms suggest that the node may have to do with datasets in the domain of culture and heritage. A word cloud of the original metadata text of datasets associated with this node is shown in Fig. 7.

Fig. 7. Word cloud for node based on metadata.

When a tag cloud is generated based on the terms in Document-Term matrix, we obtain the word cloud given in Fig. 8.

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

Fig. 8. Word cloud based on node features.

Interestingly, these features (or terms) were also found to be common when the radius of the neighbourhood for the node was increased up to 2, indicating that the entries in its neighbourhood refer to the same theme. This is further illustrated in the next section.

D. Word Clouds for Nodes & Neighbours

The low topographic error in our model is exemplified by the set of common features (or terms) shown in Table 5 for the node under consideration (radius 0) and its neighbours at radius 1 and 2.

TABLE V. COMMON FEATURES WITH NEIGHBOURING NODES

Radius	No of Nodes	No of Items	Common Features
0	1	2	`local, land, dublink, zone, area, use, dlr, plan`
1	3	5	`area, plan, dublink, land, use`
2	11	18	`dublink, use`

A consolidated Word cloud for the example node and its neighbours up to a radius of 3 is shown in Fig. 9. This confirms that datasets associated with these nodes all common terms "dublinked" and "use" are prominently shown in the map.

Fig. 9. Word cloud for Example dataset with radius of 3.

E. Evaluation

Since the datasets were not labelled ab initio, the results were presented to domain experts for evaluation. Each node and their neighbours, usually up to a radius of 2, were examined. In all cases, the experts were able to identify the topics that relate the node in question and its neighbours in the datasets. As an illustration, the example node in Table 5 is titled 'DLR Goatstown Local Area Plan' and it is labelled 121 in the map in Fig. 4 (towards the top right corner). This node, based on the two datasets it contains, was identified to refer to topics on planning and land use in DLR (Dun Laoghaire – Rathdown, Dublin). The nodes in its neighbours, up to a radius of 1, also refer to land use but without specific reference to DLR. Moving further down to the node labelled 96 (four nodes to the south west of 121). This node represents topic on planning and land use with emphasis on planning applications. The nodes that were far away on the map also clearly contained datasets on different topics. The node labelled 149 refers to topics on health and safety, while 195 refers to parks.

F. Application of Results

We briefly highlight in this section how our SOM model was employed in developing a service for recommending related dataset as part a next generation CKAN-based open data platform (Route-To-PA Platform[3]) piloted by five Local Authorities in four European countries including Republic of Ireland, Italy, the Netherlands, and France.

In the demonstrator, our SOM Dataset Recommender Service returns a list of related datasets for a given dataset. The number of datasets returned is based on the degree of relatedness specified by the user (implemented by a slider bar in the top of the list). When a user specifies high degree of relatedness, datasets that are members of the same node with the dataset of interest are returned. However, when the relatedness is relaxed, datasets associated with neighbouring nodes (within a given radius) on our SOM map are also included.

Fig. 10 and 11 are screenshots of the related dataset feature of the CKAN-based platform using our SOM-based service. Both figures show examples of related datasets returned for requests on "Luas Stops" (metro) and "Parks" in Dublin City respectively. In Figure 10, the returned results for "Luas Stops" include datasets on bus schedules, real time passenger information, bus stops for different bus operators, and traffic volumes for bridges. All these datasets are in fact very closely related to metro (Luas) stops as they all provide information that are important for commuters in Dublin City.

[3] The Route-To-PA project aims to provide Transparency Enhancing Tools (TET) as well as Social and Collaboration tools (SPOD) to extend current generation of open data platforms. See http://routetopa.eu/

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

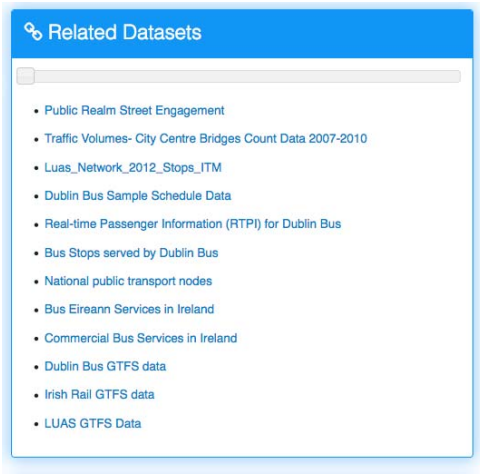

Fig. 10. Related datasets for "Luas Stops" dataset.

A more interesting example is provided Fig. 11, where the request to recommend related datasets to the "Parks" dataset produced a list of datasets on other parks, libraries, air pollution and monitoring data, trees, landscape maintenance, energy consumption. While one may initially question the notion of datasets about libraries and park as related, a closer examination reveals possible connection of the domains of these datasets. Specifically, parks, trees, energy consumption and landscape maintenance are related to "recreation and sustainable environment" domain, while in this context libraries (including mobile ones) are related to both culture and recreation. In fact, smart cities initiatives would usually treat culture and recreation as a single integrated domain as described in [23]. In addition, parks could be excellent stops for mobile libraries. Thus, integrating these related datasets has the potentials to support the design and development of smarter services.

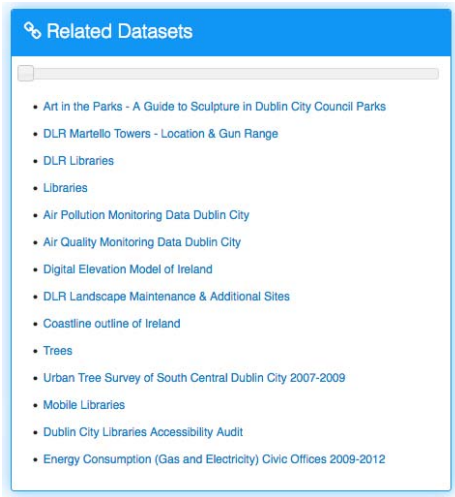

Fig. 11. Related datasets for "Parks" dataset.

V. Discussion

Open datasets pose a number of challenges, with the quality of data being one of the topmost on the list. Providing metadata information on datasets comes at a cost; one which the publishers are unwilling to foot. Thus, data quality remains an issue when it comes to open datasets. One way to address this is to take advantage of content data which inherently contains rich information and can be either textual or numeric. Textual data is easily amenable to the techniques employed in this paper, but numeric data will require a lot more novel transformations to use.

Continuous updates to dataset catalogue and increasing number of datasets published on open data platforms will necessitate update to the SOM model. The strategy for this has to be worked out in such a way that the model is up to date in a timely manner. While the model is easy to re-compute for small datasets (takes a few milliseconds for our dataset), large datasets will pose a lot more challenges. Apart from the computational requirements, large datasets will lead to large maps, thus effectively visualizing these maps will pose a challenge. An approach to handling this is to use the hierarchical SOM [24], with each hierarchy in the model giving more details of previous level. This model can also be a basis for browsing the datasets, with the ability to zoom in to finer details of a node.

Beyond relatedness, a knowledge graph can be used to represent the resulting SOM, thus providing a powerful structure to pose and answer queries, and extract valuable information on the underlying datasets.

Another promising area is to explore the social network of the datasets. Some questions we pose are: Can we view the social graph of the datasets to see how these datasets are connected to one another? Can we discover which datasets provide a link among two or more datasets or more generally the dataset with the highest centrality? Such dataset will be very interesting to discover as this may determine its value in the datasets. This can also give insights into the integration opportunities available in the different related domains and discovery of innovation opportunities.

VI. Conclusions

Our goal in this paper was to develop a SOM-based model for computing semantic relatedness among datasets in data catalogues as for recommending related datasets in open data platforms. Results provide strong evidence for the efficacy of our approach, including in revealing innovation opportunities implicit in a data catalogue. At the same time, we have noted a number challenges such as poor quality of metadata and data; that could affect the effectiveness of our approach. As part of our future work, we intend to apply our model to large scale data catalogues such as those of data.gov and data.gov.uk and integrate our model with other tools that enable integration of compatible datasets.

References

[1] Ojo, L. Porwol, M. Waqar, A. Stasiewicz, E. Osagie, M. Hogan, O. Harney, and F. A. Zeleti, "Realizing the Innovation Potentials from Open Data: Stakeholders' Perspectives on the Desired Affordances of Open Data Environment," in Collaboration in a Hyperconnected World: 17th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2016, Porto, Portugal, October 3-5, 2016, Proceedings, H. Afsarmanesh, L. M. Camarinha-Matos, and A. Lucas Soares, Eds. Cham: Springer International Publishing, 2016, pp. 48–59.

[2] World Bank, "Technical Assessment of Open Data Platforms for National Statistical Organisations," 2014.

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

[3] Lindén and J. Stråle, "AN EVALUATION OF PLATFORMS FOR OPEN GOVERNMENT DATA," Kth School of Technology and Health Handen, Sweden, 2014.

[4] Shen, Z. Riaz, M. S. Palle, Q. Jin, and F. Peña-Mora, "Open Data Platforms and Their Usability: Proposing a Framework for Evaluating Citizen Intentions," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9373, no. December 2009, pp. 247–260, 2015.

[5] Carvalho, Ç. Çalli, A. Freitas, and E. Curry, "EasyESA: A low-effort infrastructure for explicit semantic analysis," CEUR Workshop Proc., vol. 1272, pp. 177–180, 2014.

[6] Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," IJCAI Int. Jt. Conf. Artif. Intell., pp. 1606–1611, 2007.

[7] Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," Comput. Linguist., vol. 32, no. 1, pp. 13–47, 2006.

[8] A. Hassan, A. Ojo, and L. Porwol, "A Lexical Resource for Identifying Public Services Names on the Social Web," in Social Media for Government Services, S. Nepal, C. Paris, and D. Georgakopoulos, Eds. Cham: Springer International Publishing, 2015, pp. 293–324.

[9] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using Measures of Semantic Relatedness for Word Sense Disambiguation," Proc. Fourth Int. Conf. Intell. Text Process. Comput. Linguist., vol. 4, pp. 241–257, 2003.

[10] Cramer, "How well do semantic relatedness measures perform?: a meta-study," … 2008 Conf. Semant. Text Process., pp. 59–70, 2008.

[11] M. Blei, "Introduction to Probabilistic Topic Modeling," Commun. ACM, vol. 55, pp. 77–84, 2012.

[12] T. Kohonen, "The Self-Organizing Map," Proc. IEEE, vol. 78, no. 9, pp. 1464–1480, 1990.

[13] Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," IEEE Trans. Neural Networks, vol. 11, no. 3, pp. 586–600, 2000.

[14] M. Rubio and V. Giménez, "New methods for self-organising map visual analysis," in Neural Computing and Applications, 2003, vol. 12, no. 3–4, pp. 142–152.

[15] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," Proc. Work. Interact. Lang. Learn. Vis. Interfaces, pp. 63–70, 2014.

[16] World Bank, "Technical Assessment of Open Data Platforms for National Statistical Organisations," World Bank, Washingto DC, 2014.

[17] P. R. Adegboyega Ojo, Lukasz Porwol, Mohammad Waqar, Edobor Osagie, Arkadiusz Stasiewicz, Michael Hogan, Owen Harney, "Pathologies of Open Data Platform and Desired Transparency-Related Affordances for Future Platforms," 17th Annu. Int. Conf. Digit. Gov. Res., 2016.

[18] C. Alexopoulos, E. Loukis, and Y. Charalabidis, "A Platform for Closing the Open Data Feedback Loop based on Web 2 . 0 functionality," J. eDemocracy Open Gov., vol. 6, no. 1, pp. 62–68, 2014.

[19] T. Kohonen, S. Kaski, K. Lagus, J. Saloj\ärvi, J. Honkela, V. Paatero, and a. Saarela, "Self organization of a massive text document collection," Kohonen maps, vol. 11, no. 3, pp. 171–182, 1999.

[20] D. Polani, "Measures for the organization of self-organizing maps," Self-Organizing neural networks, vol. 78, pp. 13–44, 2002.

[21] J. Goodhill and T. J. Sejnowski, "Objective functions for topography: A comparison of optimal maps," 4th Neural Comput. Psychol. Work. London, 9-11 April 1997 Connect. Represent., pp. 73–83, 1997.

[22] Millar, G. Peterson, and M. Mendenhall, "Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps.," FLAIRS Conf., pp. 69–74, 2009.

[23] Caragliu, C. Del Bo, and P. Nijkamp, "Smart Cities in Europe," J. Urban Technol., vol. 18, no. 2, pp. 65–82, 2011.

[24] E. Pampalk, G. Widmer, and A. Chan, "A New Approach to Hierarchical Clustering and Structuring of Data with Self-Organizing Maps," Intell. Data Anal., vol. 8, no. 2, pp. 1–23, 2003.

978-1-5090-6435-9/17/$31.00 ©2017 IEEE